確率・統計基礎: サノフの定理、KL ダイバージェンス

森 立平

経験分布の集中

大数の法則や中心極限定理と同様に i.i.d. 確率変数 $X_1, X_2, \ldots, X_N$ を考える。

確率変数 X の像が有限集合 $A = \{1, 2, \ldots, m\}$ とする。

任意の $\mathbf{x} \in A^N$ と各 $k \in A$ について
$N_k(\mathbf{x}) := \{t \in \{1, \ldots, N\} \mid x_t = k\}$ とおく。
$\hat{P}_{\mathbf{x}} := \left(\frac{N_k(\mathbf{x})}{N}\right)_{k \in A}$ を経験分布という。

サイコロを 100 回振って 1,2,3,4,5,6 がそれぞれ 16,13,18,20,14,19 回出たとき経験分布は

$$\hat{P}_{\mathbf{x}} := \left(\frac{16}{100}, \frac{13}{100}, \frac{18}{100}, \frac{20}{100}, \frac{14}{100}, \frac{19}{100}\right).$$

経験分布

有限集合 $A = \{1, 2, \ldots, m\}$ とする。
$\mathcal{P} \coloneqq \{(p_1, \ldots, p_m) \in \mathbb{R}^m_{\geq 0} \mid p_1 + \cdots + p_m = 1\}$.

Definition

任意の $N \in \mathbb{Z}_{>0}$ について

$$\mathcal{P}_N \coloneqq \left\{ \left(\frac{N_1}{N}, \frac{N_2}{N}, \ldots, \frac{N_m}{N} \right) \,\middle|\, \begin{array}{c} (N_1, \ldots, N_m) \in \mathbb{Z}^m_{\geq 0} \\ N_1 + N_2 + \cdots + N_m = N \end{array} \right\}$$

とする。任意の $\mathbf{x} \in A^N$ について、経験分布 $\widehat{P}_{\mathbf{x}} \in \mathcal{P}_N$ は

$$\widehat{P}_{\mathbf{x}} \coloneqq \left(\frac{\widehat{N}_1(\mathbf{x})}{N}, \frac{\widehat{N}_2(\mathbf{x})}{N}, \ldots, \frac{\widehat{N}_m(\mathbf{x})}{N} \right).$$

ここで $k = 1, 2, \ldots, m$ について $\widehat{N}_k(\mathbf{x}) \coloneqq \{t \in \{1, \ldots, N\} \mid x_t = k\}$.

Example ($m = 2$ のとき)

$$\mathcal{P}_3 = \left\{ \left(\frac{0}{3}, \frac{3}{3}\right), \left(\frac{1}{3}, \frac{2}{3}\right), \left(\frac{2}{3}, \frac{1}{3}\right), \left(\frac{3}{3}, \frac{0}{3}\right) \right\}$$

$$\widehat{P}_{12211} = \left(\frac{3}{5}, \frac{2}{5}\right)$$

多項分布

Definition (カテゴリ分布)

$p \in \mathcal{P}$ について、$X \sim \mathrm{Cat}(p) \overset{\text{def}}{\iff}$

$$\Pr(X = k) = p_k \qquad \text{for } k \in \{1, 2, \dots, m\}.$$

カテゴリ分布において、X の像自体は気にしないことが多く、数ではなくてアルファベット $A, B, C, \dots$ などでもよい。

Definition (多項分布)

$p \in \mathcal{P}$ について、$X \sim \mathrm{Multinom}(p, N) \overset{\text{def}}{\iff}$

$$\Pr(X = (N_1, \dots, N_m)) = \binom{N}{N_1\ N_2\ \cdots\ N_m} \prod_{k=1}^{m} p_k^{N_k}$$

が $(N_1, \dots, N_m) \in \mathbb{Z}_{\geq 0}^m$ で $N_1 + \cdots + N_m = N$ であるものについて成り立つ。

多項分布と経験分布

Lemma

$p \in \mathcal{P}$ とする。$X_1, \ldots, X_N$ が *i.i.d.* で $\mathrm{Cat}(p)$ に従うとする。このとき、$\mathbf{X} := (X_1, X_2, \ldots, X_N)$ について $N\hat{P}_{\mathbf{X}} \sim \mathrm{Multinom}(p, N)$.

Proof.

$(N_1, \ldots, N_m) \in \mathbb{Z}_{\geq 0}^m$ で $N_1 + \cdots + N_m = N$ であるものについて

$$\Pr(N\hat{P}_{\mathbf{X}} = (N_1, \ldots, N_m)) = \binom{N}{N_1\ N_2\ \cdots\ N_m} \prod_{k=1}^{m} p_k^{N_k}$$

を示せばよい。

$$\begin{aligned}
\Pr(N\hat{P}_{\mathbf{X}} = (N_1, \ldots, N_m)) &= \Pr((\widehat{N}_1(\mathbf{X}), \ldots, \widehat{N}_m(\mathbf{X})) = (N_1, \ldots, N_m)) \\
&= \sum_{\substack{\mathbf{x} \in A^N \\ (\widehat{N}_1(\mathbf{x}), \ldots, \widehat{N}_m(\mathbf{x})) = (N_1, \ldots, N_m)}} \Pr(\mathbf{X} = \mathbf{x}) \\
&= \frac{N!}{(N_1!)\,(N_2!)\,\cdots\,(N_m!)} \prod_{k=1}^{m} p_k^{N_k}.
\end{aligned}$$

$\square$

Kullback–Leibler divergence

Definition (Kullback–Leibler divergence)

$D \colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ を以下で定義する

$$D(q \parallel p) \coloneqq \sum_{k=1}^{m} q_k \log \frac{q_k}{p_k}.$$

ただし、$0 \log 0 = 0 \log \frac{0}{0} = 0$ とし、$q \log \frac{q}{0}$ は $q > 0$ について $+\infty$ とする。

イェンセンの不等式

Lemma (イェンセンの不等式)

凸関数 $f\colon \mathbb{R} \to \mathbb{R}$ と確率変数 $X\colon \Omega \to \mathbb{R}$ について、

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

X の像が有限のときの証明.

確率変数 X は $k = 1, 2, \ldots, m$ について確率 p_k で値 a_k をとると仮定する。m についての帰納法で示す。$m = 1$ のときは明らかに成り立つ。X の像のサイズが m 未満のときにイェンセンの不等式が成り立つと仮定する。

$$\begin{aligned}
\mathbb{E}[f(X)] = \sum_{k=1}^{m} p_k f(a_k) &= \left(\sum_{k=1}^{m-1} p_k\right) \sum_{k=1}^{m-1} \frac{p_k}{\sum_{\ell=1}^{m-1} p_\ell} f(a_k) + p_m f(a_m) \\
&\geq \left(\sum_{k=1}^{m-1} p_k\right) f\left(\sum_{k=1}^{m-1} \frac{p_k}{\sum_{\ell=1}^{m-1} p_\ell} a_k\right) + p_m f(a_m) \quad \text{(帰納法の仮定)} \\
&\geq f\left(\left(\sum_{k=1}^{m-1} p_k\right) \sum_{k=1}^{m-1} \frac{p_k}{\sum_{\ell=1}^{m-1} p_\ell} a_k + p_m a_m\right) \quad (f \text{ の凸性}) \\
&\geq f\left(\sum_{k=1}^{m} p_k a_k\right) = f(\mathbb{E}[X]).
\end{aligned}$$

□

KL-divergence の性質

Lemma

任意の $p, q \in \mathcal{P}$ について $D(q \,\|\, p) \geq 0$.

Proof.

$$\begin{aligned} D(q \,\|\, p) &= \sum_{k=1}^{m} q_k \log \frac{q_k}{p_k} \\ &= -\sum_{k=1}^{m} q_k \log \frac{p_k}{q_k} \\ &\geq -\log \sum_{k=1}^{m} q_k \frac{p_k}{q_k} \qquad \text{(イェンセンの不等式)} \\ &= 0. \end{aligned}$$

□

サノフの定理

Theorem (サノフの定理)

任意の $\Gamma \subseteq \mathcal{P}$ について

$$\frac{1}{N} \log \Pr\left(\widehat{P}_{\mathbf{x}} \in \Gamma\right) \leq -\inf_{q \in \Gamma} D(q \parallel p) + \frac{m \log(N+1)}{N} \quad \textit{for any } N \in \mathbb{Z}_{>0}$$

$$\liminf_{N \to \infty} \frac{1}{N} \log \Pr\left(\widehat{P}_{\mathbf{x}} \in \Gamma\right) \geq -\inf_{q \in \Gamma^\circ} D(q \parallel p).$$

ただし、Γ° は Γ の内点の集合とする。

特定の経験分布になる確率の上下界 1/3

Lemma

任意の $q \in \mathcal{P}_N$ について

$$\frac{1}{(N+1)^m} \mathrm{e}^{-ND(q \| p)} \leq \Pr(\widehat{P}_{\mathbf{X}} = q) \leq \mathrm{e}^{-ND(q \| p)}.$$

Proof.

任意の k について $p_k > 0$ と仮定して一般性を失わない。$0^0 = 1$ とすると

$$\begin{aligned}
\Pr(\widehat{P}_{\mathbf{X}} = q) &= \binom{N}{Nq_1\, Nq_2 \cdots Nq_m} \prod_{k=1}^{m} p_k^{Nq_k} \\
&= {\color{red}\binom{N}{Nq_1\, Nq_2 \cdots Nq_m} \prod_{k=1}^{m} q_k^{Nq_k}} \prod_{k=1}^{m} \frac{p_k^{Nq_k}}{q_k^{Nq_k}} \\
&= {\color{red}\Pr(\widehat{P}_{\mathbf{Y}} = q)} \prod_{k=1}^{m} \left(\frac{p_k}{q_k}\right)^{Nq_k} \\
&= {\color{red}\Pr(\widehat{P}_{\mathbf{Y}} = q)} \mathrm{e}^{-ND(q \| p)}.
\end{aligned}$$

ここで Y は $\mathrm{Cat}(q)$ に従う確率変数。 □

特定の経験分布になる確率の上下界 2/3

Lemma

任意の $q \in \mathcal{P}_N$ について

$$\frac{1}{(N+1)^m} \mathrm{e}^{-ND(q \parallel p)} \le \Pr(\widehat{P}_{\mathbf{X}} = q) \le \mathrm{e}^{-ND(q \parallel p)}$$

Proof.

$$\Pr(\widehat{P}_{\mathbf{X}} = q) = \Pr(\widehat{P}_{\mathbf{Y}} = q) \mathrm{e}^{-ND(q \parallel p)}.$$

よって $\frac{1}{(N+1)^m} \le \Pr(\widehat{P}_{\mathbf{Y}} = q) \le 1$ を示せばよいが上界は明らか。
任意の $p' \in \mathcal{P}_N$ について

$$\Pr\left(\widehat{P}_{\mathbf{Y}} = p'\right) \le \Pr\left(\widehat{P}_{\mathbf{Y}} = q\right)$$

が成り立つと仮定する (後で証明する) と、

$$1 = \sum_{p' \in \mathcal{P}_N} \Pr\left(\widehat{P}_{\mathbf{Y}} = p'\right) \le \sum_{p' \in \mathcal{P}_N} \Pr\left(\widehat{P}_{\mathbf{Y}} = q\right) = |\mathcal{P}_N| \Pr\left(\widehat{P}_{\mathbf{Y}} = q\right)$$

であるが $|\mathcal{P}_N| = \binom{N+m-1}{m-1} \le (N+1)^m$ より下界を得る。 □

特定の経験分布になる確率の上下界 3/3

最後に

$$\Pr\left(\hat{P}_{\mathbf{Y}} = p'\right) \leq \Pr\left(\hat{P}_{\mathbf{Y}} = q\right)$$

を示せばよい。

$$\Pr(\hat{P}_{\mathbf{Y}} = p') = \binom{N}{Np'_1\, Np'_2\, \cdots\, Np'_m} \prod_{k=1}^{m} q_k^{Np'_k}$$

$$\Pr(\hat{P}_{\mathbf{Y}} = q) = \binom{N}{Nq_1\, Nq_2\, \cdots\, Nq_m} \prod_{k=1}^{m} q_k^{Nq_k}$$

より

$$\begin{aligned}
\frac{\Pr(\hat{P}_{\mathbf{Y}} = p')}{\Pr(\hat{P}_{\mathbf{Y}} = q)} &= \prod_{k=1}^{m} \frac{(Nq_k)!}{(Np'_k)!} q_k^{Np'_k - Nq_k} \\
&\leq \prod_{k=1}^{m} (Nq_k)^{Nq_k - Np'_k} q_k^{Np'_k - Nq_k} \qquad \left(\frac{b!}{a!} \leq b^{b-a}\right) \\
&= \prod_{k=1}^{m} N^{Nq_k - Np'_k} \\
&= 1.
\end{aligned}$$

サノフの定理の上界の証明

Theorem (サノフの定理)

任意の $\Gamma \subseteq \mathcal{P}$ について

$$\frac{1}{N}\log\Pr\left(\widehat{P}_{\mathbf{X}} \in \Gamma\right) \leq -\inf_{q\in\Gamma} D(q\,\|\,p) + \frac{m\log(N+1)}{N} \quad \textit{for any } N \in \mathbb{Z}_{>0}$$

$$\liminf_{N\to\infty}\frac{1}{N}\log\Pr\left(\widehat{P}_{\mathbf{X}} \in \Gamma\right) \geq -\inf_{q\in\Gamma^\circ} D(q\,\|\,p).$$

ただし、Γ° は Γ の内点の集合とする。

上界の証明.

任意の $\Gamma \subseteq \mathcal{P}$ と $N \in \mathbb{Z}_{>0}$ について

$$\begin{aligned}
\Pr(\widehat{P}_{\mathbf{X}} \in \Gamma) &= \sum_{q\in\Gamma\cap\mathcal{P}_N} \Pr(\widehat{P}_{\mathbf{X}} = q) \\
&\leq |\Gamma\cap\mathcal{P}_N| \max_{q\in\Gamma\cap\mathcal{P}_N} \Pr(\widehat{P}_{\mathbf{X}} = q) \\
&\leq (N+1)^m \max_{q\in\Gamma\cap\mathcal{P}_N} \mathrm{e}^{-ND(q\,\|\,p)} \\
&\leq (N+1)^m \sup_{q\in\Gamma} \mathrm{e}^{-ND(q\,\|\,p)} \\
&= (N+1)^m \mathrm{e}^{-N\inf_{q\in\Gamma} D(q\,\|\,p)}.
\end{aligned}$$

サノフの定理の下界の証明 1/2

Theorem (サノフの定理)

任意の $\Gamma \subseteq \mathcal{P}$ について

$$\frac{1}{N}\log\Pr\left(\widehat{P}_{\mathbf{x}} \in \Gamma\right) \leq -\inf_{q\in\Gamma} D(q \,\|\, p) + \frac{m\log(N+1)}{N} \quad \textit{for any } N \in \mathbb{Z}_{>0}$$

$$\liminf_{N\to\infty}\frac{1}{N}\log\Pr\left(\widehat{P}_{\mathbf{x}} \in \Gamma\right) \geq -\inf_{q\in\Gamma^\circ} D(q \,\|\, p).$$

ただし、Γ° は Γ の内点の集合とする。

下界の証明 1/2.

任意の $\Gamma \subseteq \mathcal{P}$ と $N \in \mathbb{Z}_{>0}$ について

$$\begin{aligned}
\Pr(\widehat{P}_{\mathbf{x}} \in \Gamma) &= \sum_{q\in\Gamma\cap\mathcal{P}_N} \Pr(\widehat{P}_{\mathbf{x}} = q) \\
&\geq \max_{q\in\Gamma\cap\mathcal{P}_N} \Pr(\widehat{P}_{\mathbf{x}} = q) \\
&\geq \max_{q\in\Gamma\cap\mathcal{P}_N} \frac{1}{(N+1)^m}\mathrm{e}^{-ND(q\,\|\,p)} \\
&\geq \frac{1}{(N+1)^m}\mathrm{e}^{-N\min_{q\in\Gamma\cap\mathcal{P}_N} D(q\,\|\,p)}.
\end{aligned}$$

サノフの定理の下界の証明 2/2

示したいことは

$$\liminf_{N\to\infty} \frac{1}{N} \log \Pr(\widehat{P}_{\mathbf{X}} \in \Gamma) \geq -\inf_{q\in\Gamma^\circ} D(q \,\|\, p).$$

今示せたことは

$$\frac{1}{N} \log \Pr(\widehat{P}_{\mathbf{X}} \in \Gamma) \geq -\min_{q\in\Gamma\cap\mathcal{P}_N} D(q \,\|\, p) - \frac{m\log(N+1)}{N}.$$

よって任意の $\Gamma \subseteq \mathcal{P}$ について以下を示せばよい。

$$\liminf_{N\to\infty} -\min_{q\in\Gamma\cap\mathcal{P}_N} D(q \,\|\, p) = -\limsup_{N\to\infty} \min_{q\in\Gamma\cap\mathcal{P}_N} D(q \,\|\, p) \geq -\inf_{q\in\Gamma^\circ} D(q \,\|\, p).$$

inf の定義より、任意の $\epsilon > 0$ について、ある $q_\epsilon \in \Gamma^\circ$ が存在して、
$D(q_\epsilon \,\|\, p) \leq \inf_{q\in\Gamma^\circ} D(q \,\|\, p) + \epsilon$.

- KL divergence の連続性より、任意の $q \in \mathcal{P}$ と $\epsilon > 0$ について、ある $\delta_{q,\epsilon}$ が存在して、$|D(q \,\|\, p) - D(q' \,\|\, p)| < \epsilon$ が任意の $q' \in S_{q,\delta_{q,\epsilon}} := \{q' \in \mathcal{P} \mid \|q - q'\| < \delta_{q,\epsilon}\}$ について成り立つ。ここで $\|\cdot\|$ はユークリッドノルム (任意のノルムでよい)。
- Γ° は開集合なので、任意の $q \in \Gamma^\circ$ についてある τ_q が存在して、$S_{q,\tau_q} \subseteq \Gamma^\circ$.

よって、$\delta := \min\left\{\delta_{q_{\frac{\epsilon}{2}},\frac{\epsilon}{2}}, \tau_{q_{\frac{\epsilon}{2}}}\right\}$ とすると、任意の $\epsilon > 0$ と $q' \in S_{q_{\frac{\epsilon}{2}},\delta} \subseteq \Gamma^\circ$ について

$$-\inf_{q\in\Gamma^\circ} D(q \,\|\, p) \leq -D\left(q_{\frac{\epsilon}{2}} \,\|\, p\right) + \frac{\epsilon}{2} \leq -D\left(q' \,\|\, p\right) + \epsilon \leq -\min_{q\in\Gamma\cap\mathcal{P}_N} D(q \,\|\, p) + \epsilon$$

が十分大きな N について成り立つ。両辺 $\liminf_{N\to\infty}$ を取ることで不等式を得る。

二項分布 $(m=2)$ の場合

二項分布の場合 $\sum_{k=1}^n X_k = N_1(\mathbf{X})$ であり、$\widehat{P}_{\mathbf{X}} = \left(\frac{N-N_1(\mathbf{X})}{N}, \frac{N_1(\mathbf{X})}{N}\right)$ と一対一に対応する ($A = \{1,2\}$ ではなくて、$A = \{0,1\}$ としている)。

$$\Pr\left(\sum_{k=1}^N X_k \geq Na\right) = \Pr\left(\frac{N_1(\mathbf{X})}{N} \geq a\right)$$

$\Gamma \coloneqq \{(p_0, p_1) \mid p_1 \geq a\}$ とすると、

$$\Pr\left(\sum_{k=1}^N X_k \geq Na\right) = \Pr\left(\widehat{P}_{\mathbf{X}} \in \Gamma\right)$$

サノフの定理より、これの指数部は

$$-\inf_{q \in \Gamma^\circ} D\left((1-q, q) \,\|\, (1-p, p)\right) = -\inf_{q \in \Gamma} D\left((1-q, q) \,\|\, (1-p, p)\right)$$

である。

$$\begin{aligned}
\inf_{q \in \Gamma} D\left((1-q, q) \,\|\, (1-p, p)\right) &= \inf_{q \geq a} (1-q)\log\frac{1-q}{1-p} + q\log\frac{q}{p} \\
&= \begin{cases} 0 & \text{if } a \leq p \\ {\color{red}(1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p}} & \text{otherwise.} \end{cases}
\end{aligned}$$

これはクラメールの定理で得られるレート関数 $I_X(a)$ と等しい。

課題

- $p = (0.1, 0.2, 0.7)$, $q = (0.3, 0.6, 0.1)$ について $D(q \| p)$ と $D(p \| q)$ をもとめよ。